会话分析视角下非正式信息交流主题与主题簇演化分析*

■ 王晓[1] 马超[2] 翟姗姗[1]
[1] 华中师范大学信息管理学院 武汉 430079 2 浙江师范大学经济与管理学院 金华 321004

摘 要：[目的/意义]针对当前非正式信息交流主题演化研究在分析层次和测度指标两方面存在的局限，提出一种具有通用性的演化分析方法，从微观和中观层面探究主题演化特征与规律。[方法/过程]引入会话分析理论，以新浪微博和知乎为例，通过对主题和主题簇运行过程进行分析，从会话内容和讨论方式两个维度揭示非正式信息交流演化特征与规律。同时，设计主题持续性计算判定方法，丰富主题演化的衡量标准。[结果/结论]主题演化分析结果显示新浪微博和知乎意见群体的发文主题存在明显偏重，且表明了意见群体参与社会焦点事件讨论中观点的主要切入角度；主题簇演化分析发现了新浪微博意见群体在一定范围内发散探索多元主题、知乎意见群体始终关注聚焦核心主题的讨论特点。两个社交媒体中意见群体在会话内容和讨论方式方面的区别，喻示了新浪微博和知乎在网络环境的非正式信息交流中主要承担的角色差异。

关键词：非正式信息交流 主题演化 主题簇 会话分析
分类号：G203
DOI：10.13266/j.issn.0252-3116.2021.17.009

1 引言

随着互联网技术发展和受到 COVID-19 疫情影响，非正式信息交流大量迁移并越发活跃于社交媒体平台中。在社交媒体中大量生成、裂变传播并迅速更新的用户生成内容（User Generated Content，UGC）里，潜藏着多种多样的主题，常被用于表征社交媒体用户的内容偏好[1]。基于社交媒体客观记录的痕迹数据，全面刻画主题演化趋势、深入探究主题演变规律，有助于准确把握非正式信息交流的特征规律，应用于具体情境中可为舆情智能监控、个性化内容推荐等提供服务参考。

现有研究中，各类主题识别方法研究成果较为丰富，为本研究的开展提供了技术保障。主题演化研究从多个方面深化拓展，但在分析层次方面相对宏观，集中于以主题为表征的特定情境中非正式信息交流的内容演化，有关主题和主题簇演化过程中内部运行情况的微观和中观层次分析较为缺乏；同时主题演化情况的测度指标着重讨论主题强度量化，忽视了从存续时间角度对主题持续性的考察。会话分析研究为非正式信息交流分析提供了社会学理论基础，但在社交媒体中的应用尚处于起步阶段。基于此，本文引入会话分析理论解析主题和主题簇演化运行过程，结合主题持续性概念内涵、测度及判定标准的界定，以期从微观和中观层面深入探究非正式信息交流演化过程及特征规律，为优化完善社交媒体平台中网络舆情管理的沟通交流策略提供参考。实证分析中，以新浪微博和知乎平台为数据来源，将 UGC 视作基于社交媒体的非正式信息交流的异步会话过程展开分析。

2 相关研究

2.1 非正式信息交流中的主题识别与演化研究

社交媒体平台为非正式信息交流中用户、互动关系及信息流的研究提供了理想环境[2]，因此网络环境下非正式信息交流研究多将新浪微博、知乎、Twitter、Facebook 及在线论坛等作为信息交流载体和实证数据来源。

主题识别研究，依据表示方法可主要分为三类：

* 本文系第 68 批中国博士后科学基金资助项目“基于社交媒体的突发事件事理图谱构建及风险预警研究”（项目编号：2020M682459）研究成果之一。
作者简介：王晓（ORCID：0000-0002-0509-7038），讲师，博士，E-mail：ray_wangx@ccnu.edu.cn；马超（ORCID：0000-0002-3425-993X），讲师，博士；翟姗姗（ORCID：0000-0002-2787-0183），副教授，博士。
收稿日期：2021-02-03 修回日期：2021-05-12 本文起止页码：91-100 本文责任编辑：易飞

①基于词的主题表示，以基于加权算法的主题识别方法为代表，将词频统计结合词性[3]、逆文档词频[4]等计算词的贡献度，通过排序筛选提取主题内容；②基于词簇的主题表示，以基于文本聚类的主题识别方法为代表，多使用 Word2vec 构建特征词集结合 K-means 聚类算法[5-6]提取文本主题；③基于概率分布的主题表示，使用主题模型识别文本主题，其中 LDA 模型因具有优秀的数据降维和隐含语义挖掘能力，被应用于多项研究中识别社交媒体短文本主题[7-8]。此外，社交媒体中大量非文本特征，如用户、地理、互动、时序等，也被分别引入主题模型[9]或结合文本内容特征构建超网络模型[10]，以实现主题联合挖掘。

主题演化研究，依据演化结构可主要分为两类，即主题线性结构演化和主题非线性结构演化。其中，前者在主题演化研究早期占主流地位[11]，主要通过主题内容或讨论强度在时间轴上的线性演变呈现，揭示以主题为表征的文本内容时序变化特征与规律。而主题非线性结构演化研究于近年来逐渐增多[12]，借助故事脉络分析[13-14]探究主题间关系的演变过程。此外，针对主题演化的发展阶段，相关研究通常以生命周期理论为基础，提出三段论[15]、四段论[16]、五段论[17]等多种划分方式。主题演化分析维度方面，不同研究中通过引入空间[18]、用户[19]等维度对单一时序加以拓展，或藉由多维特征整合[20]等方式加以丰富。

2.2 会话分析研究

会话分析理论（Conversation Analysis Theory），分属语言学和社会学两个学科领域。其中，语言学研究中的会话分析，强调语法、语篇、话轮和话题等语言形式与功能的分析[21]；社会学研究中的会话分析，则旨在通过发现人类言语交际的规律与模式阐释其背后蕴含的社会规律与社会秩序[22]。相关研究数据主要来自非正式信息交流过程中的会话记录，可分为线下会话语料和线上会话语料两类。

线下会话语料，多使用自然或半实验环境下记录人们会话交流的音频、视频，转换为文字加以整理形成。基于此的会话分析研究重点关注由序列结构分析反映出言语交际中的特征与规律，分别就外语教学[23]、医患交流[24]、跨文化工作[25]等具体情境以及协商请求[26]、故事讲述[27]等具体行为中的会话交流展开研究。

得益于互联网技术发展，网络环境中客观记录的海量交流数据推动着基于线上会话语料的会话分析研究与时俱进、日益增长。学者们从内容、关系、行为等多个维度，探究学术虚拟社区[28]、微信平台[29]等多种社交媒体中用户的信息交流特征。此外，分别对多模态大数据环境下的会话分析方法[30]以及基于会话分析的自然语言处理方法[31]、团队决策支持系统[32]等加以改进优化。

综上所述，主题识别研究成果丰富，为文本主题分析研究提供了方法和技术支持。主题演化研究从演化结构、发展阶段、分析维度等多方面得以深化拓展，但也存在一定局限：①分析层次相对宏观，集中于以主题为表征的特定情境中非正式信息交流内容演化，有关单个主题和由若干关系紧密的主题所组成的主题簇在演化过程中的内部运行情况的微观和中观层次分析较为缺乏。②主题演化情况的测度指标着重讨论主题强度量化计算，忽视了从存续时间角度对主题持续性的探究。会话分析研究提供了基于信息交流数据分析揭示人类言语交际的社会学规律的理论依据，但在基于社交媒体的非正式信息交流分析中的应用尚处于起步阶段。因此，本研究以文本主题为核心，将会话分析结合主题分析，旨在基于主题和主题簇运行过程分析，从微观和中观层面揭示非正式信息交流的特征与规律。同时，从连贯性延续和间断性延续两方面探讨主题持续性内涵，并基于相对讨论强度制定持续性判断标准，以量化分析主题演化情况。

3　基于会话分析的非正式信息交流主题演化分析框架

本文基于会话分析理论梳理主题运行过程并界定其中各类运行状态，进而结合发文数量和参与人数指标计算主题的相对讨论强度，以可视化呈现主题运行过程及衡量主题演化过程中的持续性特征。通过主题运行过程分析及其持续性测度，探究社交媒体用户在主题讨论内容中的偏重特点。

3.1　主题运行过程分析

在包含若干时间片段的时间区间中，主题由启动到终止的运行过程之间，可能出现延续、沉默、回逆等运行状态。此处有关运行状态的讨论皆限定在一定范围内，如讨论时间、讨论参与者等。主题运行过程中各种状态的具体描述如下：

（1）主题启动。指的是由一名用户发文提出新的讨论主题并介绍相关内容，可能引发其他用户参与讨论；

（2）主题延续。指的是在主题启动后，由引入该

主题的用户或其他用户接连发文，以深入挖掘或延伸扩展的方式，持续发表与该主题有关的看法；

（3）主题沉默和主题回逆。是一对相辅相成的概念，指的是主题相关发文在某一个时间片段暂时停止，但在其后处于研究选定的整体时间范围内的某一个或某几个时间片段，该主题相关发文再次出现；

（4）主题终止。指的是与这一主题有关的发文完全结束，观测范围内的所有参与者不再发表该主题相关讨论。

社交媒体平台中，不同主题可以同时被同一名用户发文讨论，同一主题可以同时被不同用户发文讨论，社交媒体中不同主题的发文多以并行关系呈现，每个主题的运行过程相对独立。但是同时，受有限时间精力的影响，用户在接收、处理与表达信息过程中具有选择性，导致不同主题之间存在竞争关系，竞争获取该时间片段中更多用户的关注与讨论。因此，探究特定时间区间中多个主题的运行过程，可通过计算各个主题的相对讨论强度，在表现主题之间此消彼长关系的同时，分析主题讨论焦点变化和判定主题持续性特征，揭示用户发文主题在内容方面的偏重及其变化。

3.2 主题相对讨论强度计算

当某一个主题占据了当前时间片段中最大比例的用户发文讨论条目时，其他主题可能不被提起，或可能被个别用户通过少量发文进行有限的讨论。若在下一个时间片段中，另一个主题取代了上一时间片段中最大比例讨论的主题，成为了这一时间片段中用户讨论的重心，关于该主题以及其他主题的讨论状态，同样存在前述两种可能。由此，计算某一主题在单个时间片段中的相对讨论强度，主要考虑围绕这一主题展开讨论的用户数量占比和发文数量占比两个指标。具体计算公式如下所示：

$$\text{Topic_Strength}(TP_{ti}^{tp}) = \left(\frac{Count(U_{ti}^{tp})}{Total_U_{ti}}\right)\cdot\alpha + \frac{Count(Post_{ti}^{tp})}{Total_Post_{ti}}\cdot\beta \qquad 公式(1)$$

其中，tp 是主题的编号，$tp=1,2,\cdots,n$；ti 是时间片段编号，$ti=1,2,\cdots,m$；$Count(U_{ti}^{tp})$ 为时间片段 ti 中参与主题 tp 讨论的用户人数，$Count(Post_{ti}^{tp})$ 为时间片段 ti 中关于主题 tp 讨论的发文数量，$Total_U_{ti}$ 为时间片段 ti 中进行发文讨论的用户总人数，$Total_Post_{ti}$ 为时间片段 ti 中的总发文数量，α 和 β 表示两个指标对于主题相对讨论强度的影响因子。TP_{ti}^{tp} 代表主题 tp 在时间片段 ti 中的相对讨论强度，取值范围为[0,1]，若 $TP_{ti}^{tp}=1$ 则表示时间片段 ti 中所有用户的所有讨论发文均为主题 tp 相关内容，若 $TP_{ti}^{tp}=0$ 则表示时间片段 ti 中没有用户发表主题 tp 相关的内容。

3.3 主题持续性及其判定

主题的持续性，表现为 UGC 中与该主题相关的讨论延续于被观察的整个时间区间，可从连贯性延续和间断性延续两种方式界定。其一，连贯性延续中，主题持续性表现为主题相关发文横跨整体时间区间中若干时间片段，即 UGC 在多个连续时间片段中均涉及该主题。其二，间断性延续中，主题持续性表现为 UGC 文本流上若干时间片段中发布与该主题相关内容条目，即该主题相关发文所存在的时间片段数量在整体时间区间中的占比超过设定阈值。上述两个角度均从时间维度出发衡量用户对某一主题的偏重，考虑用户受有限时间精力影响自发筛选可能接触和发表看法的主题内容，可通过每个时间片段中的相对讨论强度计算，实现主题的纵向演化分析和横向对比分析。

本文选择从存续时间的角度，采用连贯性延续的定义方式，结合相对讨论强度计算制定主题持续性判定标准，即在整体时间区间包含的所有时间片段中，主题的相对讨论强度均大于 0。同时设立一个例外情况，即若某一主题的相对讨论强度偶发性地为 0，仍应将该主题视为具有持续性特征。相对而言，主题的非持续性是指主题在时间区间中被短暂或频繁间断地发文讨论，即主题仅在部分时间片段中的相对讨论强度大于 0，而在其他时间片段中相对讨论强度为 0，且相对讨论强度为 0 的情况是非偶发性的。

4 基于语义关联过滤的非正式信息交流主题簇识别

由若干语义相似主题组成的主题簇，反映了社交媒体用户围绕某一事物发表各不相同但又隐性关联的观点视角，通过对主题簇的组成结构、运行过程及主导主题进行分析可综合揭示用户的讨论方式特征。其中，主题簇内部呈现出的运行状态，可揭示主题簇或稳定持续、或扩展丰富、或收敛衰退的变化过程；结合簇内各主题间持续性及相对讨论强度对比确定的主导主题，可反映用户发文中或丰富多元、或聚焦深入的讨论特征。因此，为有效提取非正式信息交流中的主题簇，本文首先分析了主题相似性常用计算方法的适用情况，其后设计关联过滤条件以确定候选相似主题对，最后探讨完整主题簇构成条件。

4.1 主题相似性计算

为衡量主题之间相似性,需首先基于 TF-IDF 计算获得表征各个主题核心内容的主题词集合,映射至语义空间得到相应主题向量以进行计算,主要指标如余弦相似度、KL 散度、对称 KL 散度、JS 散度等。

其中,余弦相似度计算方法使用主题向量之间的夹角余弦值度量其相似性,要求两个主题处于同一语义向量空间。主题 T_i 与主题 T_j 的余弦相似度,通常记为 $Sim(T_i, T_j)$,具体计算公式如公式(2)所示,余弦相似度 $Sim(T_i, T_j)$ 的值越大,主题之间的相似性就越大:

$$Sim(T_i, T_j) = \frac{|T_i| \times |T_j|}{\sqrt{T_i)^2} \times \sqrt{(T_j)^2}} \qquad \text{公式(2)}$$

若两个主题来自不同向量空间,则通常选择 KL 散度、对称 KL 散度、JS 散度等,基于主题概率分布的距离测度其相似性,以相同维度的概率分布为计算前提。即假定 p 是主题 T_i 的概率分布,q 是主题 T_j 的概率分布,p 和 q 中的概率分布维度,即词汇总数,均需为 n。若主题 T_i 和主题 T_j 的词汇空间不相同,则需对主题概率分布的来源词表进行合并增补,以保证主题概率分布中相同的维度数量和词汇项目。此三种计算方法所得结果取值越小,代表主题间概率分布的差异性越小,两个主题的相似性越大。

4.2 相似主题对选取

以相似性计算结果为基础,主题间相似关系的确定还需设定相应的过滤规则和阈值,滤除相似度低的主题对,保留真正具有相似关系的主题对,构成候选主题簇。

最简单直接的方式是选取若干对相似度较大的主题,求取它们的相似度平均值并设置为相似主题对的判定阈值。显然,这种方法受主观影响,随机性较大。若将每个主题与其他主题之间的相似性大小进行倒序排列,假设对于 T_i 而言,相似度最大的是主题 T_j,即认为 T_i 和 T_j 之间存在主题关联,则可能导致因主题之间关联关系纯粹基于最大相似度判定和构建,使得主题之间实际关联性较弱的情况产生。因此,需要采取一种改进的关联过滤方法,将主题 T_i 与主题 T_j 的语义相似性计算结果,记做 $S(T_i, T_j)$:

(1)设定一个临界阈值 ε,相似度小于该阈值的主题之间不存在相似性关联关系;

(2)对于主题 T_i 而言,与主题 T_j 和主题 T_m 皆存在大于临界阈值的相似性关联,若相似度 $S(T_i, T_m) < \theta \times S(T_i, T_j)$,θ 为设定的关联度阈值,则主题 T_i 与主题 T_j 具有更强相似性关联,相较之下主题 T_i 与主题 T_m 之间的相似性关联关系太小以至于可以被忽略不计。

经过以上过滤处理后,可将具有相似性关联关系的若干个主题对所组成的集合视为一个主题子集,即 $T_i, T_j \in TC$。

4.3 主题簇提取条件

经过关联过滤后的相似主题对,依据其间相似关系组成了若干个主题子集。主题子集需进一步通过判断条件修剪,以确保簇内主题之间具有紧密的相似关系,才能最终确定主题簇提取结果。判断主题子集是否构成一个完整的主题簇,可以考虑以下三种判断条件:

(1) 构成一个主题簇中的各个主题,两两之间皆需要存在相似性关联关系;

(2)构成一个主题簇中的各个主题,仅需要与簇内至少一个其他主题存在相似性关联关系;

(3)构成一个主题簇中的各个主题之间存在的相似关系数量,需根据簇内所包含的主题数量分情况讨论,若一个主题簇包含的主题数量大于三个,则每个主题需与主题簇内至少三个其他主题存在相似性关联关系;若一个主题簇包含的主题数量小于等于三个,则每个主题需与主题簇内其他主题两两之间存在相似性关联关系。

其中,前两种限定条件存在因过于严苛或宽泛可能导致主题簇的规模过小或过大的问题;相较而言,第三种判定条件较为恰当。因此,本文选用第三种判定条件对所得主题子集进行修剪以确定主题簇。

5 实证分析

5.1 数据采集与预处理

社交媒体中高影响力用户的发文会吸引更多追随者和普通用户的关注,由高影响力用户组成的意见群体通过呼应发文和重复曝光,引发主题关注量的指数级增长,进而影响网络舆论乃至现实事件的发展走向。鉴于此,对高影响力用户组成意见群体用户生成内容 UGC 的主题和主题簇演化运行过程进行分析并测度主题持续性,可对意见群体在讨论内容和讨论方式等方面的特点与变化进行揭示,为突发事件应对、网络舆情管理等情境中有效的沟通交流策略提供参考。

本文首先通过梳理现有研究中社交媒体用户影响力测度指标[33-35]构建用户影响力评价指标体系,采用层次分析法确定指标权重,分别识别新浪微博和知乎平台中参与社会焦点事件“江歌案”讨论的高影响力

用户组成社交媒体平台的意见群体样本,采集两组意见群体为期 37 个月的发文作为本研究实证分析数据来源。其后,清洗发文数据,获得新浪微博意见群体共计 124 556 条有效发文数据,知乎意见群体共计 2 833 条有效发文数据。经过分词、去停用词等预处理后,调用百度 AI 的词向量表示功能并结合 TF-IDF 计算筛选每篇发文的特征词组成文本向量,使用 Python 中的 Scikit-learn 函数计算误差平方和与轮廓系数确定最优主题数量 K 值,通过 K_means 聚类得到 70 个新浪微博意见群体主题和 35 个知乎意见群体主题。最后,计算各主题类团中关键词的 TF-IDF 值并将降序排名前 10 词语作为主题词描述主题内容,并参考新浪微博与知乎以及清博指数、知微事见等舆情网站和人民网、新浪新闻等新闻网站确定主题标签,以方便后续分析表述。

5.2 非正式信息交流主题演化分析

首先,依据知乎和新浪微博平台来源实验数据的时间分布特点,分别选择适合的时间片段划分单元。然后,依据主题相对讨论强度计算公式(公式 1),计算每个主题在各个时间片段中的相对讨论强度。最后,基于主题相对讨论强度,对选定实验时间跨度范围内新浪微博和知乎平台中用户讨论主题的运行情况及其持续性展开分析。

就在线主题讨论而言,由多名用户发文讨论点的主题,通常比仅被少数用户讨论的主题更为活跃、传播范围更广。因此,主题相对讨论强度计算公式(公式(1))的两个影响因子 α 和 β 中,参与用户指标对于主题相对讨论强度的影响显然要大于相关发文指标,同时参考已有研究中的相关设定[29],令 $\alpha = 0.6$, $\beta = 0.4$。同时,本文将主题相对讨论强度为 0 的"偶发性"界定为 1 次,即若某一主题的相对讨论强度仅在 1 个时间片段中取值为 0,则其仍被视为具有持续性。

(1)新浪微博中的主题演化分析。新浪微博中发文较为频繁,因此本文分别考虑以天和以周为单位时间片段进行划分。若以天为单位时间片段,则依据本文判定标准,新浪微博意见群体主题均不具备持续性特征,且时间片段缺失数量较多。因此,本文选择以周为单位时间片段进行新浪微博意见群体的主题演化分析,为期 37 个月的整体时间区间被划分为 161 个时间片段,后文中以"TS + 编号"表示。指标参数经统计代入相对讨论强度计算公式(公式(1))中,得到新浪微博中各个主题在各时间片段中的相对讨论强度。以 Topic 2、Topic 36 和 Topic 47 作为示例,绘制主题运行过程,如图 1 所示:

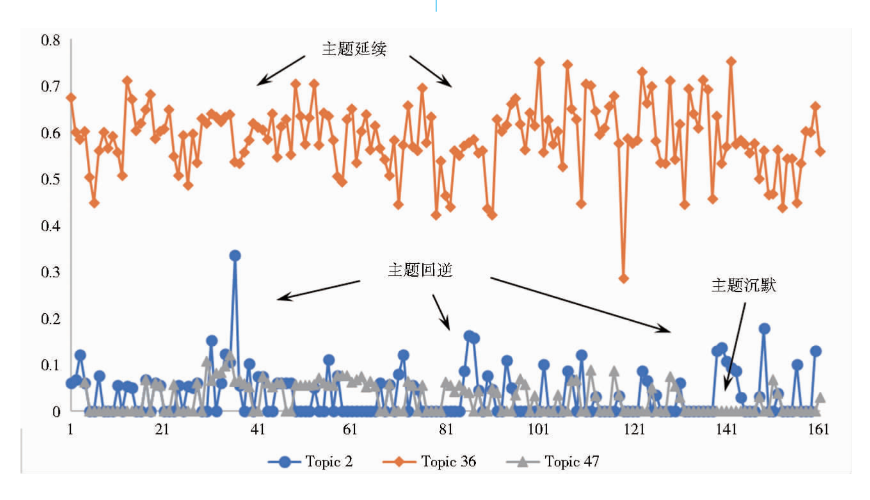


图 1　新浪微博意见群体主题运行过程示例

由图 1 可知,内容以观点评论为主的 Topic36 在整个时间区间内具有较高的相对讨论强度,表现出明显的持续性特征,反映了新浪微博意见群体对个人观点意见具有持久、强烈的表达欲望。Topic2 和 Topic47 的相对讨论强度整体较低,且在若干个时间片段中出现主题沉默,相对讨论强度非偶发性地取值为 0,不具有持续性特征。其中,Topic2 围绕着节日红包展开,周期性的主题回逆表现突出;Topic47 主要与方舟子、彭剑诈骗安保资金案有关,作为由高影响力用户主导和推动吸引其他网民关注的舆情事件,其相对讨论强度的分布变化呈现出较为完整的生命周期过程。即 Topic47 相关的发文从曝光相关信息(TS4)开始,随着事件进展接连披露相关信息并延伸相应讨论;在 TS30 至 TS39 和 TS49 至 TS65 期间达到主题讨论的高潮阶段,其间主题相对讨论强度较大且时间片段相对连续;此后该主题发文呈现频繁的沉默与回逆状态,表示相关

讨论进入衰退期；最后在 TS132 及其后进入沉默状态，直至相关讨论彻底结束，主题终止。

（2）知乎平台中的主题演化分析。基于知乎平台发文时间间隔特点，本文考虑了以周或月为单位时间片段的划分方案。若以周为单位时间片段，由于知乎用户发文的时间分布较为稀疏，知乎意见群体主题均不具备持续性特征。因此，需按月进行划分，共计 37 个时间片段，以“TS + 编号”表示。相关参数数值经统计代入公式 1 中，计算知乎意见群体主题在各时间片段中的相对讨论强度。以 Topic 1、Topic7、Topic16 和 Topic 17 作为示例，绘制主题运行过程，如图 2 所示：

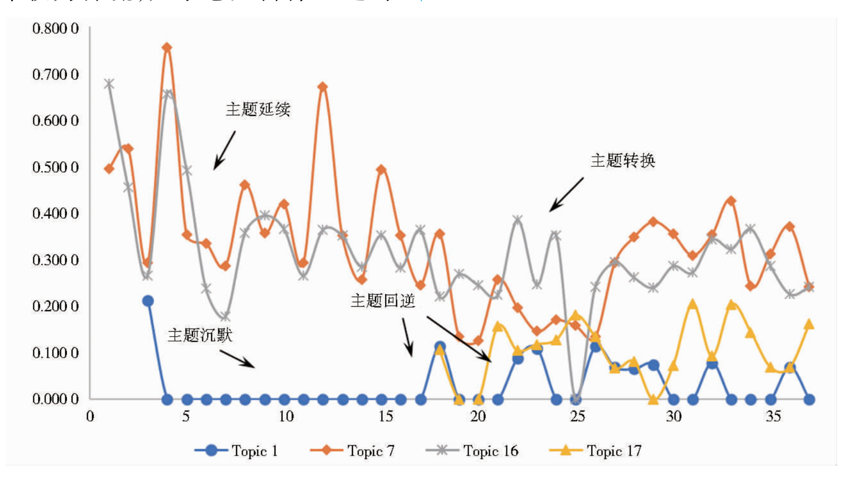


图 2　知乎意见群体主题运行过程示例

图 2 中可以看出，Topic7 在启动后的所有时间片段中相对讨论强度均大于 0，具有明显的持续性特征，且相较于其他三个主题更加被意见群体发文所偏重，反映知乎意见群体从法律角度进行专业性解读的特点，并将此偏好特点延续至社会焦点事件的讨论中。Topic16 自启动后，除在 TS25 中相对讨论强度偶发性地为 0，其余所有时间片段中均被意见群体发文讨论，也具有持续性特征，表明受知乎平台对发文内容详尽解析的鼓励，意见群体将突发性社会焦点事件与个人长期兴趣爱好（如文学作品）相结合以表达自身观点的特点。同时，意见群体自 TS18 开始发文参与社会民生主题（Topic17）相关讨论，尽管在初期（TS19 和 TS20）出现短暂沉默，但后续时间片段中的相对讨论强度较为平稳，显示出该意见群体逐步发展出对社会事件的讨论参与偏重。此外，Topic1 在多个时间片段中的沉默状态和在 TS18 及其后时间片段中频繁的沉默与回逆交替状态，均显示出电影娱乐等无关主题不是基于社会焦点事件讨论形成的意见群体所共享的集体偏好，该类主题仅代表个别成员的兴趣爱好且通常不会与事件信息进行关联的隐喻表达。

依据本文制定标准判定，新浪微博意见群体发文中共计 6 个主题具有持续性特征，相应主题内容反映出该平台中由社会焦点事件讨论形成的意见群体对于社会事件、社会名人动态与访问等社会类主题以及对政务权威信息发布、观点意见与情感交流等主题的参与偏好。知乎中共计 3 个主题具有持续性特征，揭示出该平台意见群体参与社会焦点事件的讨论是基于对海外留学等事件相关内容主题的日常关注，并在意见表达时结合法律等专业特长和文学等兴趣爱好进行解读的交流内容特征。

5.3　非正式信息交流主题簇演化分析

主题相似性计算中，因本研究中主题向量来源于同一语义空间采用余弦相似度衡量主题相似性。其次关联过滤中，临界阈值 ε 的取值范围参考现有研究设定为 0.2 – 0.4；关联度阈值 θ 的取值范围，相关研究中通常被设定在 0.5 – 0.7 之间[36]，本文实验分别计算了新浪微博和知乎中 $\theta = 0.7$、$\theta = 0.65$ 和 $\theta = 0.6$ 三种取值方案所提取的候选相似主题对数量，依据可提取的候选主题簇数量与规模确定新浪微博中 $\theta = 0.65$ 和知乎中 $\theta = 0.7$，得到新浪微博意见群体发文中 48 对相似主题和知乎中 41 对相似主题，通过可视化工具 Gephi 绘制主题的相似关系网络，见图 3。

由图 3 可以看出，新浪微博意见群体主题相似关系中，Topic3 与 Topic42 之间存在明显相似关系，且二者分别与 Topic66 存在较强相似性，其余各个主题之间的相似关系较弱且交织复杂。知乎意见群体主题相似关系中，Topic2 与 Topic22 之间存在明显语义相似性，Topic3 与 Topic26、Topic10 与 Topic25、Topic11 与 Topic25 之间

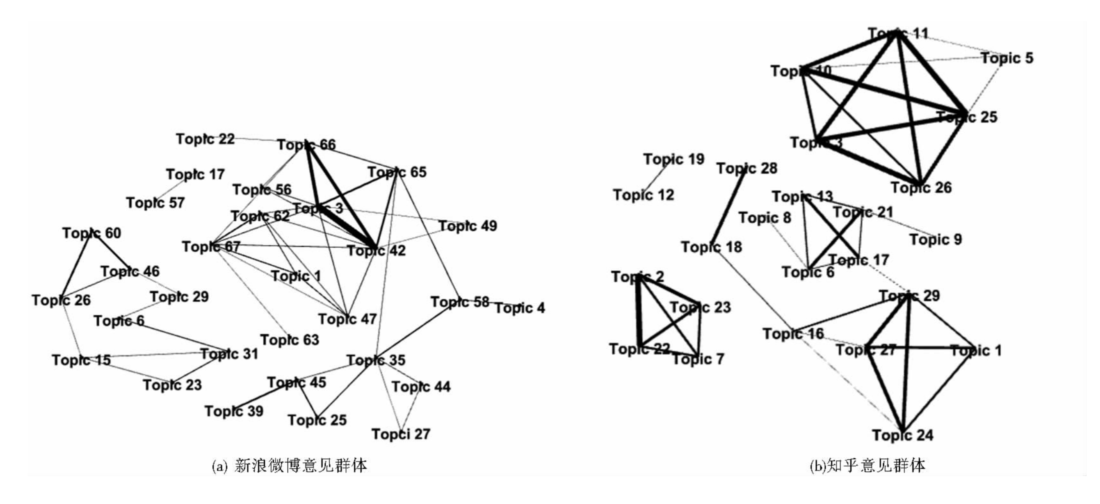


图 3　主题相似关系网络

均具有较强的相似性，各个主题自然而然地形成了若干个较为明显的主题子集。

(1)新浪微博中的主题簇演化分析。依据 4.3 小节中讨论选用的主题簇提取条件，本文在新浪微博意见群体的 70 个发文主题中共计提取 14 个主题簇，其中仅 1 个主题簇由 3 个以上相似主题构成，其余 5 个主题簇分别由 3 个相似主题构成，8 个主题簇分别由 2 个相似主题构成。构成主题簇的不同主题，体现出意见群体围绕同一议题从不同视角出发进行分析解读与意见表达，揭示了不同主题之间在 UGC 蕴含的主观认识中建立的隐性关联，有助于对意见群体发文讨论的切入角度进行更加丰富且深入的理解。例如，主题簇 1 共包含了 9 个主题，涉及社会、生活、政务、时政、突发事件等多类主题，显示出新浪微博意见群体对与人民群众日常生活及利益息息相关的多个方面内容的关注；构成主题簇 2 的 3 个主题主要涉及政务和时政类内容，反映了新浪微博意见群体在对政府工作动态的关注中，延伸出对于国内城市建设与管理、国际交流与合作等的讨论。以主题簇 2 为例，绘制其中主题在各时间片段中相对讨论强度的变化，如图 4 所示，分析主题簇内运行过程中各主题之间关系。



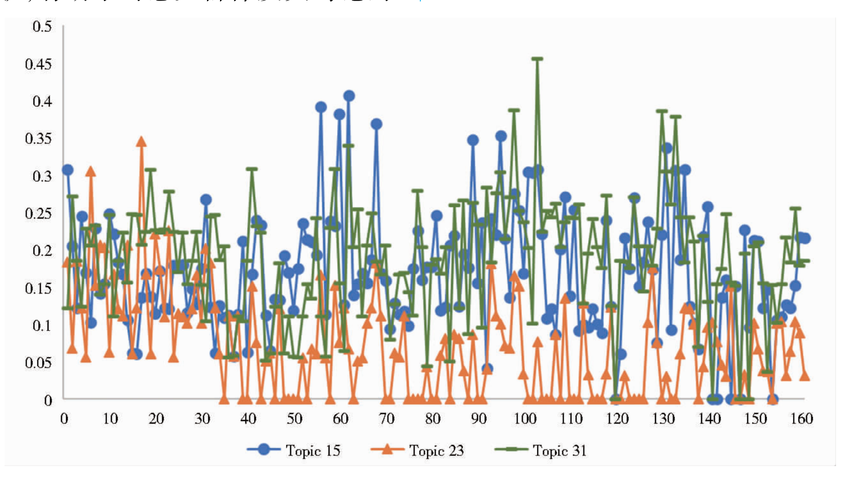


图 4　新浪微博中主题簇 2 运行过程示例

由于新浪微博中时间片段较多且该主题簇内各主题的讨论强度变化频繁，导致图 4 中散点折线的交织复杂，因此未在图中标识出各种运行状态。可以看出，其中 Topic15 和 Topic31 在多数时间片段中呈现延续状态，且根据相对讨论强度值的变化显示出两个主题交替占据意见群体在该主题簇内的讨论侧重；Topic23 则在一段时间的频繁发文（TS1 至 TS32）之后逐渐消退，主题运行在沉默与回逆两种状态之间频繁转换。在分析所有新浪微博主题簇的运行过程和主导主题后，可以发现绝大多数主题簇与图 4 中主题簇 2 的运行状态相似，其中由多个活跃主题交替主导主题簇内讨论揭示出新浪微博主题簇演化过程中内容丰富、多元的特点。此外，少数仅包含 2 个主题的主题簇呈现出由 1 个主题占据绝大部分讨论的运行状态，即主题

簇 11、12 和 14，其运行过程与 5.2 小节图 1 中所示相似。

（2）知乎平台中的主题簇演化分析。依据相同主题簇提取条件，从知乎意见群体的 35 个发文主题中共计提取 11 个主题簇，其中共 4 个主题簇分别由 3 个以上相似主题构成，其余 7 个主题簇分别由 2 个相似主题构成。

构成主题簇的不同主题同样体现出知乎意见群体从不同视角分析同一议题的解读与表达特点，但相较于新浪微博意见群体发文主题簇涵盖内容较为丰富、呈现出横向关联特征而言，知乎意见群体发文中提取的主题簇在涵盖内容范围上相对聚焦、呈现出纵向深入特征。以包含 3 个以上主题的主题簇为例，主题簇 1 主要涵盖影视娱乐内容，主题簇 2 专注法律专业内容，主题 3 聚焦竞技体育内容，主题簇 4 偏重社会生活内容。这些主题簇中，部分内容与意见群体参与社会焦点事件讨论中选择的剖析视角相关联，其他内容则揭示出意见群体具有较为广泛且持续的兴趣爱好。以主题簇 1 为例，绘制其中各个主题的相对讨论强度的变化，如图 5 所示：

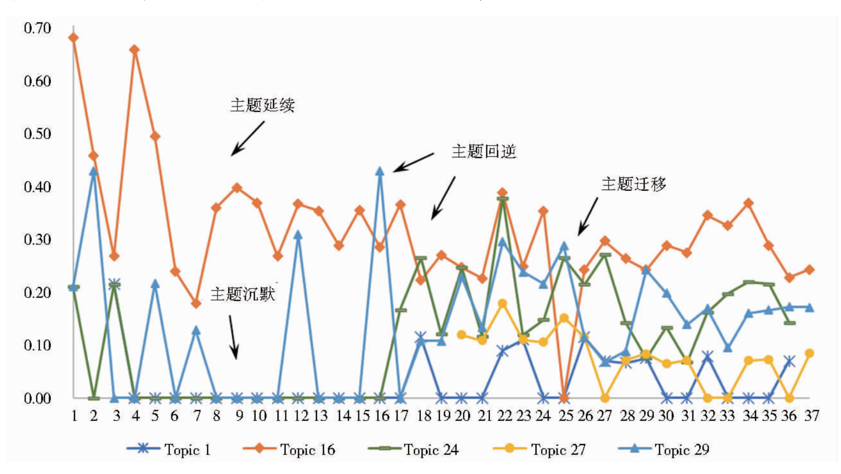


图 5　知乎中主题簇 1 运行过程示例

从图 5 中可以看出，意见群体前期（TS1 至 TS17）讨论以 Topic16 为主，间或谈及其他主题，整体内容较为单一；而在后期（TS17 之后）发文中，不同主题的发文增多，该主题簇中的讨论内容更加丰富；主题簇涵盖内容随时间推移逐渐拓展丰富的这一特点，在知乎意见群体发文数据中较为普遍。具体运行过程中，则发现 Topic16 不仅具有持续性特征，且在多数时间片段中占据了所述主题簇讨论的热点地位。其余主题的运行过程中，均存在或短或长时间的沉默期和沉默与回逆之间频繁的状态切换。尽管 Topic24 和 Topic29 在 TS17 之后呈现出围绕较高水平的相对讨论强度均线进行波动的状态，且在个别时间片段中居于主题簇 1 的热点地位，但在多数情况下 Topic16 仍是主题簇 1 中的核心主题。在分析所有知乎主题簇的运行过程和主题簇内主导主题后，可以发现各知乎主题簇均分别由一个主题占据主导地位，簇内其余主题或存在较长时间的沉默期、或在沉默与回逆间频繁切换，整体而言体现出知乎意见群体专注聚焦核心主题的讨论方式特点。

6　总结与展望

本文引入会话分析理论，通过分析主题和主题簇运行过程揭示非正式信息交流的微观和中观层面的演化特征与规律，并提出主题持续性探索演化分析衡量标准。实证分析中以社会焦点事件中高影响力用户构成的意见群体为例，分析新浪微博和知乎中的主题和主题簇运行过程，揭示意见群体在非正式信息交流的主题讨论内容和方式上的偏重特点与变化趋势，以期为网络舆情管理中制定有效沟通交流策略提供参考。

分析发现，主题持续性反映了意见群体在该主题内容上的明显偏重，并表明了意见群体在社会焦点事件讨论中意见观点的主要切入角度。同时，新浪微博和知乎意见群体在具有持续性特征的主题之间存在的明显差异，揭示出两个社交媒体平台有关社会焦点事件讨论中高影响力用户在事件相关和日常状态的 UGC 中主题内容角度的区别，表示二者在网络环境的非正式信息交流中承担的角色差异。由候选相似主题对集合形成的关系网络展现了新浪微博中内容交织复杂、边界模糊与知乎中内容相似性差异明显、边界清晰的

特点,源于两个平台中 UGC 发文特点、主题识别方法等多重因素影响,在此情况下,主题簇提取条件中采用相似关系数量判定的方法有助于完整主题簇的准确判定。同时,主题簇运行过程分析,展现了新浪微博意见群体在一定范围内发散探索不同主题,知乎意见群体始终关注聚焦核心主题的讨论特点。

本文研究还存在一定局限与不足,后续研究中可从以下两个方面进行完善:其一,从间断性延续的角度探讨主题持续性,并与本文研究中的连贯性延续进行对比,以进一步丰富主题演化衡量标准研究;其二,从主题和主题簇演化运行角度对比分析事件相关与常规状态中 UGC 内容的异同,与评价对象与情感倾向角度相结合,丰富有关突发事件刺激情境下认知相符与失调研究的分析层次。此外,还可从挖掘用户在跨平台 UGC 中意见表达变化关联性的角度进行拓展,以加深对网络环境下非正式信息交流中用户意见表达的认识。

参考文献:

[1] 王臻皇, 陈思明, 袁晓如. 面向微博主题的可视分析研究[J]. 软件学报, 2018, 29(4): 1115－1130.
[2] BEX R T, LUNDGREN L, CRIPPEN K J. Scientific Twitter: the flow of paleontological communication across a topic network[J/OL]. PLoS one, 2019, 14(7). [2021－05－06]. http://doi.org/10.1371/journal.pone.0219688.
[3] BOKAETF M H, SAMETI H, LIU Y. Unsupervised approach to extract summary keywords in meeting domain[C]// DUGELAY J L, SLOCK D. Proceedings of the 23rd European signal processing conference. Piscataway: IEEE, 2015: 1406－1410.
[4] CHEN Y H, LU J L, MENG F T. Finding keywords in blogs: efficient keyword extraction in blog mining via user behaviors [J]. Expert systems with applications, 2014, 41(2):663－670.
[5] 谷莹, 李贺, 李叶叶, 等. 基于在线评论的企业竞争情报需求挖掘研究[J]. 现代情报, 2021, 41(1): 24－31.
[6] 安璐, 李倩. 基于热点主题识别的突发事件次生衍生事件探测[J]. 情报资料工作, 2020, 41(6): 26－35.
[7] ZHANG Y H, MAO W J, ZENG D, et al. Topic evolution modeling in social media short texts based on recurrent semantic dependent CRP[C]// BENJAMIN V, LI W F. Proceedings of 2017 IEEE international conference on intelligence and security informatics. Piscataway: IEEE. 2017: 119－124.
[8] 廖海涵, 王曰芬, 关鹏. 微博舆情传播周期中不同传播者的主题挖掘与观点识别[J]. 图书情报工作, 2018, 62(19): 77－85.
[9] 梁晓贺, 田儒雅, 吴蕾, 等. 微博主题发现研究方法述评[J]. 图书情报工作, 2017, 61(14): 141－148.
[10] 梁晓贺, 田儒雅, 吴蕾, 等. 基于超网络的微博相似度及其在微博舆情主题发现中的应用[J]. 图书情报工作, 2020, 64(11): 77－86.
[11] SASAKI K, YISHIKAWA T, FURUHASHI T. Online topic model for Twitter considering dynamics of user Interests and topic trends [C]// MARTON Y. Proceedings of 2014 conference on empirical methods in natural language processing. Stroudsburg: ACL,2014: 1977－1985.
[12] LIU Y P, PENG H, LI J X, et al. Event detection and evolution in multi-lingual social streams[J/OL]. Frontiers of computer science, 2020, 14(5). [2021－05－06]. http://doi.org/10.1007/s11704-019-8201-6.
[13] DEHGHANI N, ASADPOUR M. SGSG: semantic graph-based storyline generation in Twitter[J]. Journal of information science, 2019, 45(3): 304－321.
[14] GOYAL P, KAUSHIK P, GUPTA P, et al. Multilevel event detection, storyline generation, and summarization for Tweet streams [J]. IEEE transactions on computational social systems, 2020, 7(1): 8－23.
[15] HUANG J J, PENG M, WANG H, et al. A probabilistic method for emerging topic tracking in microblog stream[J]. World Wide Web, 2017, 20(2): 325－350.
[16] CAI H Y, HUANG Z, SRIVASTAVA D, et al. Indexing evolving events from Tweet streams[J]. IEEE transactions on knowledge and data engineering, 2015, 27(11): 3001－3015.
[17] ABULAISH M, FAZIL M. Modeling topic evolution in Twitter: an embedding-based approach [J/OL]. IEEE access, 2018, 6. [2021－05－06]. http://doi.org/10.1109/ACCESS.2018.2878494.
[18] PRUSS D, FUJINUMA Y, DAUGHTON AR, et al. Zika discourse in the Americas: a multilingual topic analysis of Twitter[J/OL]. Plos one, 2019, 14(5). [2021－05－06]. http://doi.org/10.1371/journal.pone.0216922.
[19] 王臻皇, 陈思明, 袁晓如. 面向微博主题的可视分析研究[J]. 软件学报, 2018, 29(4): 1115－1130.
[20] 刘雅姝, 张海涛, 徐海玲, 等. 多维特征融合的网络舆情突发事件演化话题图谱研究[J]. 情报学报, 2019, 38(8): 798－806.
[21] SACKS H, SCHEGLOFF E A, JEFFERSON G. Simplest systematics for the organization of turn-talking for conversation[J]. Language, 1974, 50(4): 696－735.
[22] 吴亚欣, 于国栋. 为会话分析正名[J]. 山西大学学报(哲学社会科学版), 2017, 40(1): 85－90.
[23] 赵焱, 张旗伟, 徐蕊, 等. 超语及认同建构作为双语者的学习手段[J]. 现代外语, 2021(2): 258－270.
[24] STOMMEL W, VAN GOOR H, SYOMMEL M. Other-attentiveness in video consultation openings: a conversation analysis of video-mediated versus face-to-face consultations[J]. Journal of computer-mediated communication, 2019, 24(6): 275－292.
[25] AVISON D, BANKS P. Cross-cultural (mis)communication in IS

offshoring: understanding through conversation analysis[J]. Journal of information technology, 2008, 23(4): 249－268.
[26] 吴亚欣，刘蜀. 请求行为之微妙性的序列组织研究[J]. 现代外语，2020，43(1)：32－43.
[27] 彭欣，张惟. 日常交谈中故事讲述的会话分析[J]. 山西大学学报(哲学社会科学版)，2019，42(4)：137－144.
[28] 卢恒，张向先，张莉曼，等. 会话分析视角下虚拟学术社区用户交互行为特征研究[J]. 图书情报工作，2020，64(13)：80－89.
[29] 巴志超，李纲，毛进，等. 微信群内部信息交流的网络结构、行为及其演化分析——基于会话分析视角[J]. 情报学报，2018，37(10)：1009－1021.
[30] GU Y, LI X Y, HUANG K X, et al. Human conversation analysis using attentive multimodal networks with hierarchical encoder-decoder[C]// ACM. Proceedings of the 26th ACM multimedia conference. New York: ACM, 2018: 537－545.
[31] HOUSLEY W, ALBERT S, STOKOE E. Natural action processing: conversation analysis and big interactional data[C]// ACM. Proceedings of the halfway to the future symposium. New York: ACM, 2019: 1－4.
[32] KONO S, AIHARA K. Prototype of decision support based on estimation of group status using conversation analysis [C]// YAMAMOTO S. Proceedings of the 18th international conference on human-computer interaction. Berlin: Springer, 2016: 40－49.
[33] 张星，魏淑芬，王莉，等. 危机事件中的微博意见领袖影响因素实证研究[J]. 情报学报，2015，34(1)：66－75.
[34] CUI L, PI D C. Identification of micro-blog opinion leaders based on user features and outbreak nodes [J]. International journal of emerging technologies in learning, 2017, 12(1): 141－154.
[35] 安璐，胡俊阳，李纲. 突发事件情境下社交媒体高影响力用户画像研究[J]. 情报资料工作，2020，41(6)：6－16.
[36] 郭晓利，周自岚，刘耀伟，等. 基于 DTS-ILDA 模型和关联过滤的新闻话题演化分析[J]. 应用科学学报，2017，35(5)：634－646.

作者贡献说明：

王晓：提出研究思路与研究框架，收集分析数据，撰写论文；
马超：处理分析数据；
翟姗姗：提出论文修改意见，参与论文修订。

Evolutionary Analysis of Topic and Topic Clusters in Informal Communication from the Perspective of Conversation Analysis

Wang Xiao[1] Ma Chao[2] Zhai Shanshan[1]
[1] School of Information Management, Central China Normal University, Wuhan 430079
[2] College of Economic and Management, Zhejiang Normal University, Jinhua 321004

Abstract: [**Purpose/significance**] Aiming at the limitations of current informal communication topic evolution research in both analysis level and measurement indicators, a universal evolution analysis method is proposed to explore the characteristics and patterns of topic evolution from micro and medium levels. [**Method/process**] Introducing the conversation analysis theory, taking Sina Microblog and Zhihu as examples, this paper revealed the evolutionary characteristics and patterns of informal information communication from the two dimensions of conversation content and discussion style through the analysis of running process of topics and topic clusters. Meanwhile, this paper designed the method of calculating and judging the continuity of a topic and explored measurement standard of the topic evolution. [**Result/conclusion**] The topic evolution analysis results show that the opinion group from Sina Microblog and Zhihu are obviously biased in topic content, and indicate the main perspectives of opinion group participating in the discussion of social focus event. The topic cluster evolution analysis find out that opinion group from Sina Microblog diversify and explore multiple topics in a certain range, while those from Zhihu always focus on the settled core topics. The difference in conversation content and discussion style between opinion groups in social media indicates the different role of Sina Microblog and Zhihu in informal information communication online.

Keywords: informal communication topic evolution topic cluster conversational analysis